SEARCHING FOR PHONOLOGICAL AMELIORATION

Joshua Fennell

A research paper submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Linguistics in the College of Arts and Sciences.

Chapel Hill
2021

Approved by:

Katya Pertsova

Elliott Moreton

ABSTRACT

Joshua Fennell: Searching for Phonological Amelioration
(Under the direction of Katya Pertsova)

Many models of phonotactics predict that the judged acceptability of words should generally decline as word length increases. However, it has been observed that in some cases, a word containing elements of poor acceptability may have its overall acceptability ameliorated by other elements of better acceptability. This paper discusses an experiment performed to test the hypothesis that a word of poor acceptability can see its acceptability improve by appending a string of better acceptability, and secondarily, the hypothesis that acceptability indeed declines with length. The results of the experiment show no evidence to support the first hypothesis, but some degree of support for the second.

ACKNOWLEDGEMENTS

Here I wish to express my sincere gratitude to Dr. Katya Pertsova for the patient and steadfast support and guidance she provided over the course of this research project and throughout my time at UNC.

This research was also supported in part by an NSF grant, “Inside Phonological Learning” (Award Number 1651105), to UNC-Chapel Hill (PI Elliott Moreton).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

It is known that speakers of a language have an innate intuition that allows them to judge the phonotactic acceptability ("well-formedness", "wordlikeness") of strings of sounds as possible words of their language. Speakers do so in a gradient fashion, judging words to be not simply absolutely acceptable or unacceptable, but somewhere along a continuum in between. The hypothetical mental system providing this capability is termed a *phonotactic grammar*. It is argued by some (Daland 2015) that such a grammar should be probabilistic in nature, in that it would associate a probability distribution with the set of all possible strings formed from the sounds of the language, with the grammaticality of a word derived from its probability (measurable acceptability judgments being assumed to reflect grammaticality as computed by such a grammar). General models of phonology are also often probabilistic, assigning probability distributions to output forms; such models include the Partially Ordered Constraints model, Stochastic OT, and Maximum Entropy grammars (Coetzee and Pater 2011).

An objection to the conflation of grammaticality with probability raised by Heinz (Pater 2015) is as follows: if grammaticality is equivalent to probability, it should be possible, paradoxically, for a very long but relatively well-formed string to have a probability (and thus grammaticality, in spite of being well-formed) lower than a short, ill-formed string (a "too-much-of-a-good-thing" situation). The objection rests upon the nature of probabilistic models that under such models the probability of strings generally decreases as they increase in length. One approach to resolving this problem is to propose that the overall grammaticality of a string be in some fashion derived from the grammaticality of smaller local parts of the whole, thus avoiding the property of decline in grammaticality with length.

In particular, one phenomenon captured by the "grammaticality of local parts" type of model but not typical probabilistic models is *phonological amelioration*. This is the possibility for phonotactic violations in one part of a word to be offset by the presence in the word of phonological material (i.e., syllables or other sequences of segments) of greater acceptability.

Coleman and Pierrehumbert (1997) reported the observation of exactly this phenomenon in a set of acceptability judgments on nonsense words.

This project was carried out to further investigate the reality of the phenomenon of phonological amelioration. An experiment was conducted wherein acceptability judgments of English-like nonsense words were collected for the purpose of detecting this effect, and secondarily, testing the hypothesis emerging from the predictions of probabilistic phonotactic models that words should decline in acceptability with increasing length. The collected acceptability judgments were specifically compared against competing predictions of two different reference phonotactic models, one of which follows the approach described above of deriving grammaticality from local parts of a string by averaging bigram probabilities.

The two hypotheses tested, each one supported by one of the two reference models, are as follows:

1. Amelioration Hypothesis: It is possible for shorter words of poor acceptability to be "improved" through the addition of extra phonological material of better grammaticality.
2. Length Hypothesis: All other things being equal, words will decline in acceptability as they increase in length.

2 Experiment overview

To test the two experimental hypotheses, an experiment ("Amelio") was conducted wherein participants provided acceptability judgments on nonsense word stimuli in three groups:

- poorly acceptable ("Bad") monosyllables
- highly acceptable ("Good") disyllables
- trisyllables consisting of concatenations of Good and Bad components

The primary purpose of the experiment was to determine whether the Good-plus-Bad

concatenations would receive higher ratings than the Bad words on their own — evidence of amelioration.

To conduct the experiment, a set of known Good and Bad nonsense words was required. The process of generating candidate stimuli is described in Section 3. Pre-final stimuli were selected based on scores assigned by two reference phonotactic models which make contrasting predictions regarding amelioration. These words were then tested in a micro-experiment ("Premelio") as a means of checking the suitability of the stimuli for use in Amelio proper. The two reference phonotactic models are described in the following sections.

2.1 The Averaging Bigrams ("AvBg") model

The first reference model is a version of the widely known n-gram model of grammaticality. Under this sort of model, the grammaticality of a word is computed in some fashion based on the probabilities of the n-grams (overlapping sub-sequences of segments of length n) that comprise the word. (Albright 2008)

For example, assuming n-grams of length 2 (*bigrams*), the nonsense word /snɔtʃ/ has five: #/s/, /sn/, /nɔ/, /ɔtʃ/, /tʃ/#. (Word boundary symbols may also be included at the beginning and end of a word, as is the case here, represented by the symbol #.)

Each bigram has an associated transitional probability, computed according to the formula shown in Figure 1.

$$p(s_k \mid s_{k-1}) = \frac{lexicon\_count(s_{k-1}s_k)}{\sum_s lexicon\_count(s_{k-1}s)}$$

Figure 1: Formula for computing transitional bigram probability

In this formula, s_k represents a given segment, and $lexicon\_count()$ computes the total number of occurrences of a bigram within a lexicon.

Two possible approaches to combining individual n-gram probabilities to compute overall grammaticality in an n-gram model are:

1. the product of n-gram probabilities
2. the average of n-gram probabilities

Under either approach, the greater the score, the greater the grammaticality of the word.

As a reference phonotactic model supporting the Amelioration Hypothesis, a model that computes overall grammaticality by averaging bigram probabilities was adopted. Averaging was chosen over multiplication of probabilities because of the clear predictions it makes regarding the Amelioration Hypothesis: that amelioration can occur, and that there is no general tendency for words to become less grammatical simply due to an increase in length. Averaging of bigram probabilities has also been used to assess phonotactic probability in prior work, e.g., Vitevitch et al. (1997).

Given the use of averaging, overall grammaticality of a word is computed as shown in Figure 2.

$$g = \frac{\sum_b p(b)}{N}$$

Figure 2: Formula to compute overall word grammaticality for the AvBg model

In this formula, b represents a given bigram in a word, and N is the total number of bigrams in the word.

Bigram probabilities for this model are taken from an estimate of their relative frequency within the English lexicon. (As in the example above, word boundary symbols are included at the beginning and end of a word.) The frequency of bigrams within the lexicon (i.e., set of word types) is preferred over frequency within speech in accordance with the finding in Albright (2008) that token frequency provides no advantage to this sort of grammatical model (or to the others

examined in that paper).

2.2 The Maximum Entropy Phonotactics ("MaxEnt") model

The second reference model is the Maximum Entropy model of phonotactics proposed by Hayes and Wilson (2008). Under this model, the grammaticality of a word is computed as the weighted sum of the number of violations incurred by the word of each of a set of phonotactic constraints. The constraints themselves take the form of weighted prohibitions against specific sets of n-grams (where the sets are described using natural classes based on a feature set provided to the learning algorithm). Formally, the grammaticality of a word is computed as shown in Figure 3.

$$g = w_1 v_1 + ... + w_n v_n$$

Figure 3: Formula to compute overall word grammaticality for the MaxEnt model

In this formula, w_i represents the weight of constraint i and v_i is the count of violations of constraint i in a word.

It is important to note that in contrast to the grammaticality scores under the Averaging Bigrams model, where higher scores indicate greater grammaticality, MaxEnt model scores measure the degree to which a word is penalized by the grammar, and thus increase with decreasing word grammaticality. Also in contrast to the AvBg model, MaxEnt predicts that amelioration should not in general be observed; violation of constraints within a word is additive, and portions of a word that do not introduce violations cannot "balance out" violations in other parts of the word.

Despite the general predictions made regarding amelioration by the reference models, the use of string concatenation to test the Amelioration Hypothesis introduces a complication, which is discussed in the following section.

2.3 Concatenation and prediction of amelioration

The basic predictions of the reference models regarding amelioration become more complicated when concatenation of Good and Bad strings is introduced. The act of concatenating strings produces a result string that is not identical in its phonological content to the sum of the original strings. Specifically, the word boundaries at the site of the concatenation are eliminated, and a new juxtaposition between segments in the original strings is created. This has consequences regarding whether or not the reference models may predict amelioration.

For the AvBg model, concatenation results in the loss of two bigrams containing word boundary symbols where the two strings are joined, and the addition of new "hinge" bigram, containing no word boundary, in their place. For this reason, the AvBg score for the new, concatenated word will not be exactly identical to the average of the probabilities of all bigrams in the original words. The exact effect of this change will vary depending on the strings that were concatenated. In particular, if the hinge bigram is of sufficiently low probability, or if the Good string loses much of its Goodness due to loss of its word boundary bigram, it could offset the gains from the more acceptable material in the Good string. Thus, the model will not always predict amelioration in situations where it might be expected.

The MaxEnt model sees a similar complication, for the opposite reason. Loss of the word boundary in the Bad string has the potential to eliminate constraint violations. If the Good string introduces minor or no new violations, amelioration could occur simply due to the loss of the word boundary. The characteristic of the model to accumulate penalties still means that Good parts of a word cannot obscure violations in Bad parts, but concatenation as performed in this experiment means that amelioration is not strictly impossible.

3 Stimuli

The stimuli consisted of a larger set of mono- and disyllables which were pre-screened in the Premelio micro-experiment, from which a subset was chosen for Amelio based on participant ratings. This initial set consisted of 56 monosyllables and 56 disyllables (each of these sets further divided into 28 intended to be Good and 28 intended to be Bad).

It was desirable that all stimuli be presented as the same part of speech in order to be subject to the same phonotactic restrictions. For this reason, all stimuli were presented as nouns. These pseudo-nouns were further designed to be no less acceptable than the least-acceptable existing English noun, the intent being to avoid "features that are unrepresentative of the whole range … to which the results should generalize" (Schutze 2017).

3.1 Raw material

In order to generate stimuli resembling English nouns, existing nouns were needed as "raw material" for construction as well as statistics on various properties of English nouns. For this purpose, all mono-morphemic noun lemmas were extracted from the CELEX (Baayen et al. 1996) database. Mono-morphemic lemmas were chosen in order to avoid unusual consonant clusters that may arise at, e.g., compound boundaries, and to avoid an over-representation in the word set of plural and other affixes. This yielded a set of 4,871 nouns. For use in stimulus generation, lists of all attested syllable onsets, nuclei, and codas were additionally extracted from this set of nouns.

For logistical reasons concerning both the speaker of the stimuli and expected participant pool, American English pronunciations were preferred over the British contained in CELEX. Thus, pronunciations were extracted from the CMU Pronouncing Dictionary (Carnegie Mellon University). The final set of nouns consisted of only mono- and disyllables with pronunciations in CMUdict, 3,544 items.

3.2 Candidate pools

The extracted sets of syllable onsets, nuclei, and codas were used to generate maximal sets of mono- and disyllables, according to separate segmental patterns for each group. The patterns used were as follows (where O and C represent an onset or coda of a single segment, OO and CC two segments, and N represents a nucleus):

- monosyllables: ON, OON, ONC, OONC, ONCC, OONCC
- disyllables: ON-ON, OON-ON, ON-OON, ONC-ON

Notably, monosyllables were permitted to have simple or complex codas, whereas disyllables were not allowed word-final codas. These choices were made to allow Bad monosyllables greater opportunity to violate phonotactic constraints, and reduce such opportunity for Good disyllables.

Additionally, in an attempt to control for lexical neighborhood influence on judgments, all candidate stimuli were filtered to have a Levenshtein distance (Navarro 2001) of at least 2 with the set of CMUdict pronunciations for *all* English noun lemmas extracted from CELEX (NB: a larger set than the extracted set of nouns described above).

3.3 Stress

It was noted that allowing only a single stress pattern for the concatenated trisyllables to be used in Amelio would considerably limit the vowels available for certain positions in the disyllables that would begin each trisyllable (and ultimately occur as standalone stimuli). For this reason, it was decided to use two common English stress patterns for the trisyllable: primary-unstressed-secondary (“1-0-2”), and unstressed-primary-unstressed (“0-1-0”). Since in all cases, concatenation was performed with the disyllabic component first, all disyllables would then have a stress pattern of either primary-unstressed (“1-0”) or unstressed-primary (“0-1”).

Each stress pattern has its own restrictions on vowels that may appear in given word

positions in existing English nouns. Accordingly, in constructing the stimuli, 1-0 disyllables were permitted to have almost any vowel in their first syllable, but only either [ə] or [i] in their second. In contrast, 0-1 disyllables were only permitted to have [ə] or [ɪ] in their first syllable, but could end in [eɪ], [u], [aɪ], [ɑ], [ɔ], [aʊ], or [ɔɪ]. (Some vowels that commonly appear with primary stress as the middle vowel of a trisyllabic noun are not permitted to appear word-finally in English: [æ], [ɛ], [ɪ], [ʊ].)

3.4 Selection of final stimuli

The segmental patterns used in generation of monosyllabic and disyllabic candidates were chosen to make monosyllables and disyllables more *likely* to be Good and Bad, respectively, but it was still necessary to select stimuli from each group to be presented to the Premelio participants as Good or Bad. This task required scoring the candidate stimuli for their predicted acceptability, and the two reference phonotactic models described above were used for this purpose.[1]

To determine a reference range of acceptability for the stimuli, the set of training nouns was first scored using both models. Selected stimuli were not permitted to have scores worse than the least-acceptable training noun scores for each model.

Good and Bad stimuli were selected using a standard of "consensus" between the two reference models: only words that the models agreed were Good or Bad were selected for either level of acceptability. To qualify as Bad, a word needed to sustain violations of at least one MaxEnt model constraint (only around 13% of the training nouns have such violations), and have an AvBg score less than the 10th percentile of training noun scores. To qualify as Good, a word was required to have no violations of MaxEnt model constraints (a score of 0), and an AvBg score

[1] Here it must be noted that an undetected error in the AvBg model scoring software resulted in incorrectly computed scores during the stimulus generation phase. Thanks to the Premelio screening process, this error did not affect the use of the selected stimuli in the main experiment: human participants had the final say in the stimulus selection process. Unfortunately, however, many of the scores computed by the corrected AvBg model do not conform to the requirements described in this section, although it was still possible to use the corrected scores in a correlation analysis, discussed in Section 5.5.3.

greater than the 25th percentile of training noun scores. (The training nouns are all existing nouns of English, and by virtue of this should mostly be considered Good, a fact reflected in their MaxEnt scores.)

Final stimuli were selected to meet these scoring criteria and to produce sufficient phonological diversity within the stimuli as a whole to avoid having any one stimulus sound too much like another.

3.5 Recording of audio stimuli

The audio stimuli were recorded by the author using the Praat (Boersma and Weenink 2021) software application and a CAD D189 dynamic microphone with wind screen and pop filter in an enclosed, echo-dampened space. Stimuli were recorded at a sampling rate of 44.1 kHz with a single audio channel. To prevent any difference in recorded audio quality due to, e.g., speaker fatigue (Schutze 2017), stimuli were recorded in blocks of four, consisting of one word from each of the groups of stimuli (warm-up stimuli were recorded separately). For consistent intonation across stimuli, each word was recorded within the frame sentence, “Show me the X”. All stimuli were recorded at least three times each, with the best selected for actual use.

4 Pre-screening micro-experiment (“Premelio”)

As the final step in the stimuli selection process, the Premelio micro-experiment was conducted to ensure that the stimuli participants in Amelio, the main experiment, would be exposed to met the required criteria of being Good or Bad. The sole purpose of Premelio was the screening of stimuli; it was conducted at small scale, and no other significant data analysis was attempted.

4.1 Participants

The experiment was conducted online, with participants recruited via Amazon Mechanical Turk. Participants were presented with a pre-questionnaire requesting the following pieces of information:

- general age range
- highest completed level of education
- native language
- other languages spoken to any degree
- description of any prior experience with or knowledge of linguistics

Potential participants were also informed that in order to participate, they must have no significant hearing concerns, and were required to complete a sound check.

Data was collected from a total of 30 participants. 16 were excluded in accordance with the exclusion criteria described below, leaving a total of 14 participants:

- Listed something other than English as their native language.
- Suspected of repeating the experiment multiple times.
- Produced obviously fake ratings (e.g., all one value or simply alternating values).
- Took an excessive amount of time and/or took breaks.
- Repeatedly assigned the highest rating to words with flagrant phonotactic violations.

4.2 Design

Stimuli consisted of a total of 116 nonsense words of one or two syllables in length, falling into five groups:

- 28 Bad monosyllables (“Bad1”)

- 28 Good disyllables (“Good2”)
- 28 Good monosyllable distractors (“Good1”)
- 28 Bad disyllable distractors (“Bad2”)
- 4 warm-up stimuli, each conforming to one of the above types

All participants in the experiment were exposed to the same stimuli, in different random orders. The warm-up stimuli were presented in a separate group before beginning the main part of the experiment.

Before beginning the warm-up section, participants were shown a screen containing the following instructions:

> On each screen, you will click a button and listen to a **meaningless nonsense word**. You will then decide, based on how the word sounds, how natural it would be if it were a **noun** of English, and give it a rating from **1 to 5 stars**. (A **noun** is a name for something; for example, **“dog”**, **“air”**, **“happiness”**, and **“work”** are all nouns.)
>
> For example, you could even give star ratings to actual English words based on how natural you think they sound. Your ratings might be similar to those shown below:
>
> - genre ★
> - curfew ★★★
> - bundle ★★★★★
> - vertex ★★
>
> It is this sort of feeling you will rely on when making judgments about the nonsense words you hear. You don’t need to try and analyze the words in any particular way; just assign ratings quickly based on how you feel. There are no right or wrong answers!
>
> The spelling will also be shown for each word to help clarify the sounds you hear. Try and base your judgments only on the sound of the word, not the spelling.

During the rating task, participants were shown the following prompt: “Assign a star rating based on how natural you think the word sounds as a noun of English, from one star (completely bizarre) to five stars (completely normal).” (adapted from Albright and Hayes (2003))

Stimuli were presented in both audio and written form. A button was present onscreen allowing participants to replay the audio as many times as desired. There was no time limit to complete an individual rating task, although the main rating section as a whole was required to be completed within one half-hour. All stimuli were prefixed with the definite article “the” in both audio and written form. (For the audio, the same recording of the definite article was used for all stimuli.)

Stimulus written forms were chosen to be consistent and as unambiguous as possible. For more on stimulus written forms, see APPENDIX B: ORTHOGRAPHY.

4.3 Results

Overall, participants produced ratings in keeping with general expectations: most stimuli intended to be Good were rated better than those intended to be Bad. Figure 4 shows a box plot of the mean ratings for each stimulus across participants by group.

4.3.1 Stimulus selection problem

The highest- and lowest-rated 24 monosyllables and disyllables were selected to form the Good and Bad stimulus groups for Amelio (based on the mean ratings for each stimulus). However, this resulted in a situation where the highest mean rating of the new Bad monosyllable group was identical to the lowest mean rating of the new Good disyllable group (2.79). It was decided that this was unsatisfactory, given that these groups of stimuli were intended to have opposite acceptability polarities.

To address this situation, the pool of candidate stimuli was expanded using stimuli included

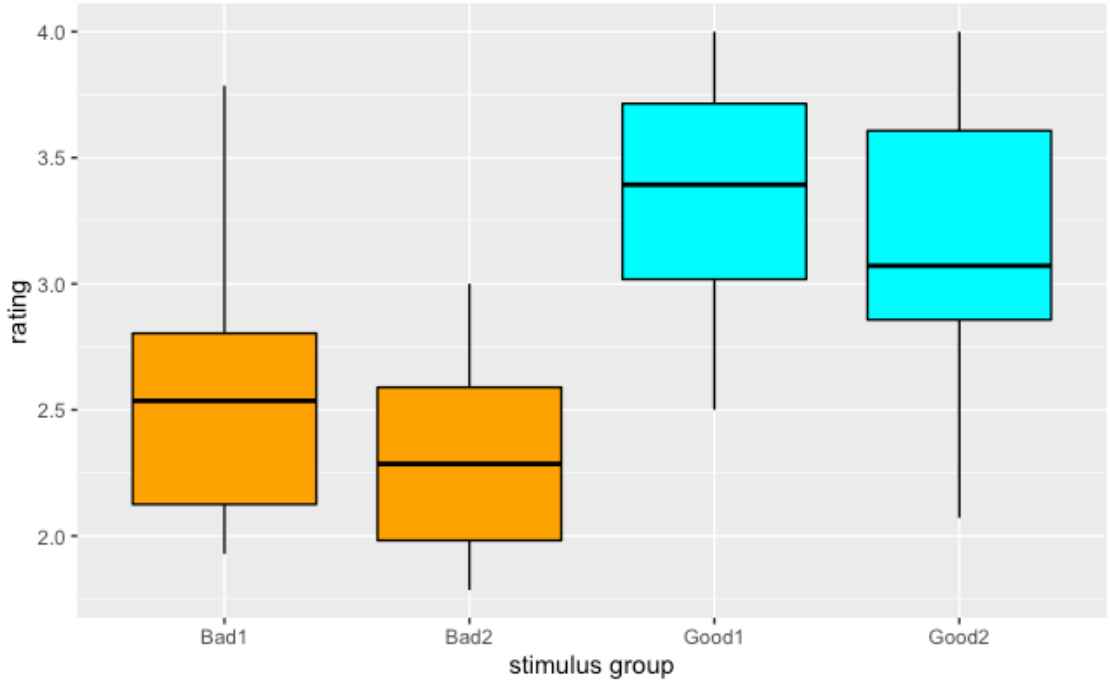


Figure 4: Premelio stimulus mean ratings by group

in a prior, discarded version of Premelio. This initial version of the experiment was conducted using disyllables with word-final codas, and a set of monosyllables that only partially overlapped with those used in the final version of Premelio. As a solution to the problem described above, all monosyllables (Good and Bad) from the earlier version of Premelio were included as candidate Amelio stimuli. With this inclusion, the new sets of highest- and lowest-rated monosyllables prevented any overlap in the ranges of group item ratings.

5 Main experiment ("Amelio")

This section describes the structure and results of the main experiment conducted for this project, dubbed "Amelio". The primary purpose of this experiment was to test the Amelioration Hypothesis, which states that the phonological acceptability of a word known to be considered of poor acceptability can be improved by concatenating the word with another word known to be of better acceptability. Results of the experiment were also used to evaluate a secondary hypothesis,

the Length Hypothesis, which states that in general, the acceptability of words tends to decrease as they increase in length.

5.1 Design

Figure 5 gives a visual overview of the structure of the experiment.

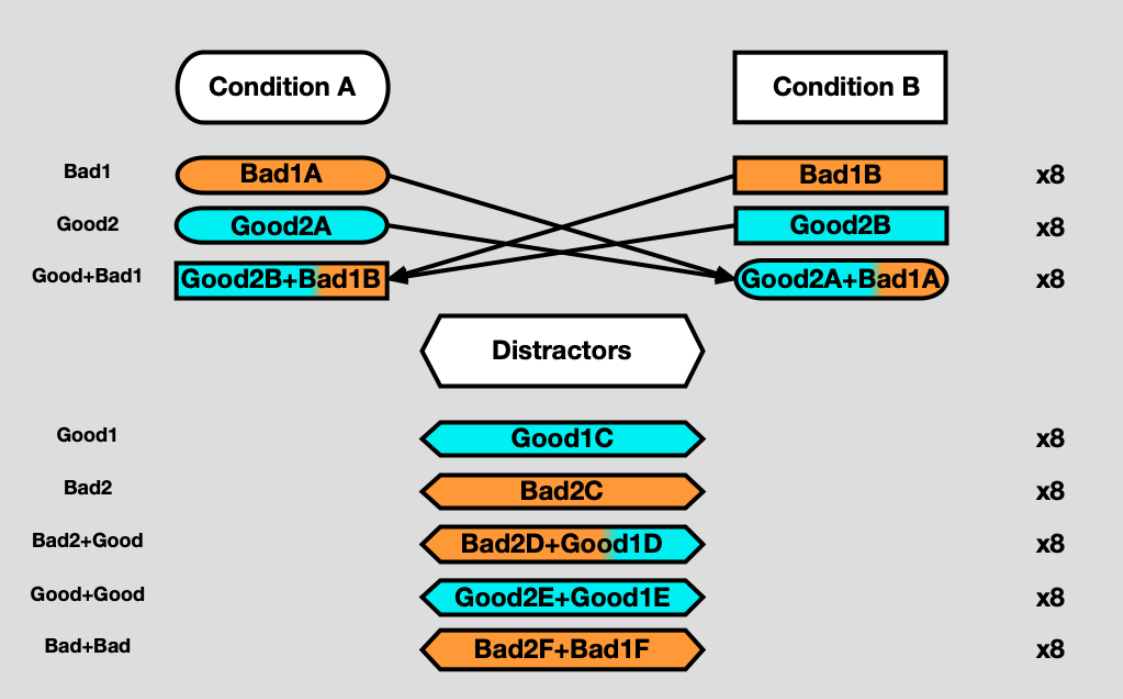


Figure 5: Visual overview of Amelio structure

As seen in Figure 5, participants in Amelio were placed into one of two conditions: A and B. Each condition was associated with a different set of eight Bad monosyllabic and eight Good disyllabic stimuli to which participants were exposed, represented in the figure by the boxes labeled Bad1A, Good2A, Bad1B, and Good2B. Participants in each condition were also exposed to a third set of "Good+Bad1" trisyllabic stimuli. These stimuli were constructed by selecting a unique element from the Good2 group of the opposing condition and appending an element of the Bad1 group from the same (i.e., opposing) condition. In this fashion, all members of the Good+Bad1 group consisted of the concatenation of a member of the Bad1 stimulus group

concatenated with a member of the Good2 group (in the order Good2 + Bad1). This allowed for a direct comparison of the ratings of the Bad1 group and the Good+Bad1 group to test the Amelioration Hypothesis. Forming the Good+Bad1 group stimuli for one condition from standalone stimuli from the opposing condition ensured that participants in a given condition would not be exposed to any stimulus both in a standalone fashion and as part of a concatenated stimulus.

All participants, regardless of condition, were also exposed to the same set of "distractor" stimuli in five other groups of eight: Good monosyllables, Bad disyllables, and Good+Good, Bad+Bad, and Bad2+Good trisyllables. As shown in the figure, the trisyllable distractors were constructed from separate mono- and disyllables that did not occur as standalone stimuli; as with the Good+Bad1 group, all of these trisyllables consisted of an initial disyllable concatenated with a final monosyllable. The distractor stimuli primarily served the purpose of attempting to mask patterns that might otherwise be perceived in the main stimuli (e.g., "all monosyllables are Bad"). However, ratings collected for the distractors would also prove useful for hypothesis testing.

5.2 Concatenated stimuli

The Premelio screening micro-experiment produced groups of 24 Bad monosyllables, Good disyllables, Good monosyllables, and Bad disyllables for use in Amelio. A total of five groups of concatenated trisyllables were constructed from these parts, with the remaining mono- and disyllables occurring as standalone stimuli.

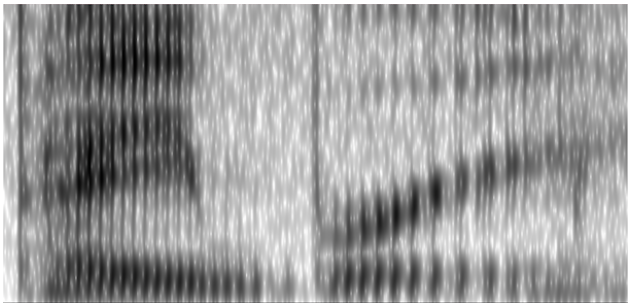
Recording the required trisyllabic Good+Bad1 stimuli as whole units separately from the mono- and disyllables they are comprised of would have inevitably introduced audible differences in various phonetic qualities between the trisyllables and their component standalone stimuli. Such differences would have complicated comparison between the Bad1 and Good+Bad1 stimulus ratings. In order to avoid this issue, the Good+Bad1 audio stimuli for a given condition were constructed by concatenating the audio for one Good disyllable and Bad monosyllable from

the opposing condition. However, simple concatenation of the unmodified audio files of the standalone stimuli would have produced unnatural results in terms of intonation, and an incorrect result in terms of stress. For this reason, systematic edits to the standalone stimuli were required before concatenation, and the edited versions of the standalone stimuli were used in the final Bad1 and Good2 groups to avoid the phonetic difference problem described above. The trisyllabic distractors were constructed using the same methodology, although in this case, the component mono- and disyllables did not occur as independent stimuli.

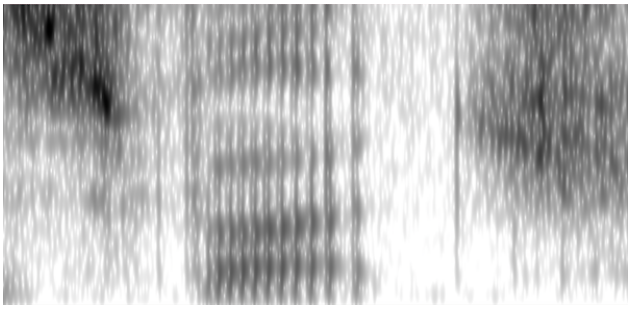
5.2.1 Concatenation method

The concatenation process described below was performed entirely using Praat. The steps were as follows:

1. In order to produce components for the concatenation that would not sound overly long when concatenated, as much audio as possible was deleted from the end of each mono- and disyllable, with the requirement that the result remain intelligible as the intended sequence of sounds. This at times resulted in the edited standalone stimuli sounding somewhat clipped, but produced an overall more natural timing of the syllables in the concatenation.
2. To avoid popping in the final audio, the ends of both components were faded out to silence over 5 ms, and the start of the second (i.e., monosyllabic) component was faded in from silence for 5ms.
3. The Praat “Scale peak” function was used to scale the peak intensity of the disyllable to 0.2, and the monosyllable to 0.075 for 1-0-2 concatenations and 0.1 for 0-1-0 concatenations.
4. The Praat “Change gender” function was used to set a new pitch median of 110 Hz for the first (i.e., disyllabic) component and 75 Hz for the second component. This step and the previous were intended to produce a realistic stress pattern in the final word.
5. The two edited components were concatenated.



(a) paibry



(b) sphutch

Figure 6: Spectrograms of raw stimuli audio

Figure 6 shows spectrograms of the raw, unedited audio for Condition A Good2 and Bad1 stimuli ‘paibry’ [ˈpeɪbɹi] and ‘sphutch’ [sfʌtʃ], respectively, before the process described above has been applied. Figure 7 shows a spectrogram of the end result of applying this process to produce the Condition B Good+Bad1 stimulus ‘paibrysphutch’.

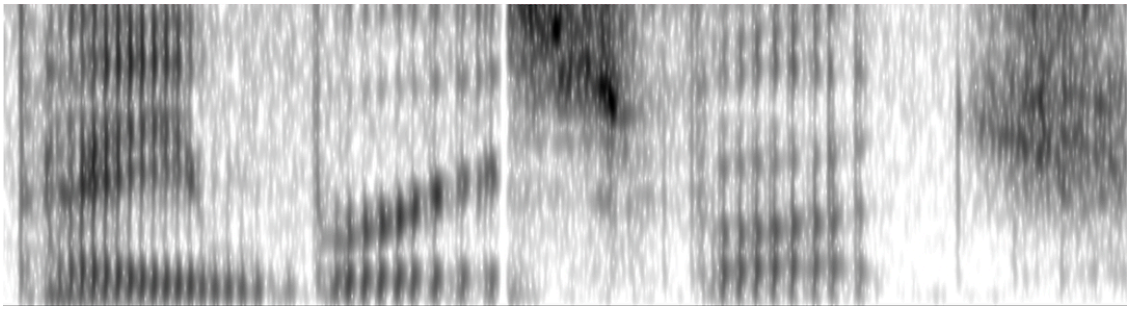


Figure 7: Spectrogram of concatenated stimulus ‘paibrysphutch’

This process is somewhat crude, and produced mixed results. It is possible that concatenated stimuli produced using this method could be judged poorly based on their sometimes unnatural sound, although a separate experiment could be conducted to test this hypothesis and possibly correct for any such effect (discussed in Section 7).

5.3 Other differences from Premelio

This section describes a number of minor changes in experiment format between Premelio and Amelio.

5.3.1 Modifications to instructions

The instructions given to participants were changed to omit mention of the stimuli being nouns. Results from one participant indicated that the emphasis on the noun status of stimuli could lead to stimuli being rated strictly on their "noun-ness" (i.e., similarity to a prototypical noun) vs., e.g., "adjectivity" or "adverbiality" (similarity to other specific parts of speech) as opposed to phonological acceptability. After this change, the noun status of stimuli was conveyed more incidentally by the prepending of the definite article.

Additionally, a sentence was added to the instructions warning participants that the onscreen "spellings" may look "a bit strange", and to please only treat them as a guide to the sounds heard in the audio. This change was made in a further effort to dissuade participants from relying on the orthography when making judgments.

5.3.2 Priming

Participants in Premelio were shown existing English nouns with suggested star ratings as a means of "calibrating" their internal rating scales. These were replaced in Amelio with disyllabic nonsense words (in audio form with spelling). One word was presented for each possible star rating. All priming words were disyllabic to prevent association of syllable length with star rating. Participants were not, however, required to play the audio of these example words. The new priming words were as follows:

- "the zherpny" ★ (['ʒɜɾpni])

- “the kigba” ★★★ (['kɪgbə])
- “the crulloo” ★★★★★ ([kɹə'lu])
- “the baiska” ★★★★ (['beɪskə])
- “the vuchay” ★★ ([və'tʃeɪ])

5.3.3 Exit question

Following completion of the main rating section of the experiment, participants were asked to “describe the process you used when making judgments about the nonsense words in as much detail as possible”.

5.4 Participants

Participants were recruited in the same manner as Premelio. The identical pre-questionnaire was administered, with the exception that the majority of participants did not receive the native language question due to a programming error. This was rectified for a minority of participants toward the end of the run of the experiment.

A total of 137 participants took part in the experiment. 37 were excluded due to violations of the exclusion criteria described in Section 4.1, leaving data from 100 participants.

5.5 Results and analysis

Figure 8 shows a box plot of mean stimulus ratings across participants by group. Table 1 shows the overall mean rating for all stimuli in each group. As seen in the plot and table, trisyllabic stimuli (right four groups), regardless of group, were typically rated notably lower than the mono- and disyllables, even those intended to be Bad. However, ratings within the trisyllabic groups appear generally sensible, with the Good+Good group having the highest ratings and the

Bad+Bad group the lowest.

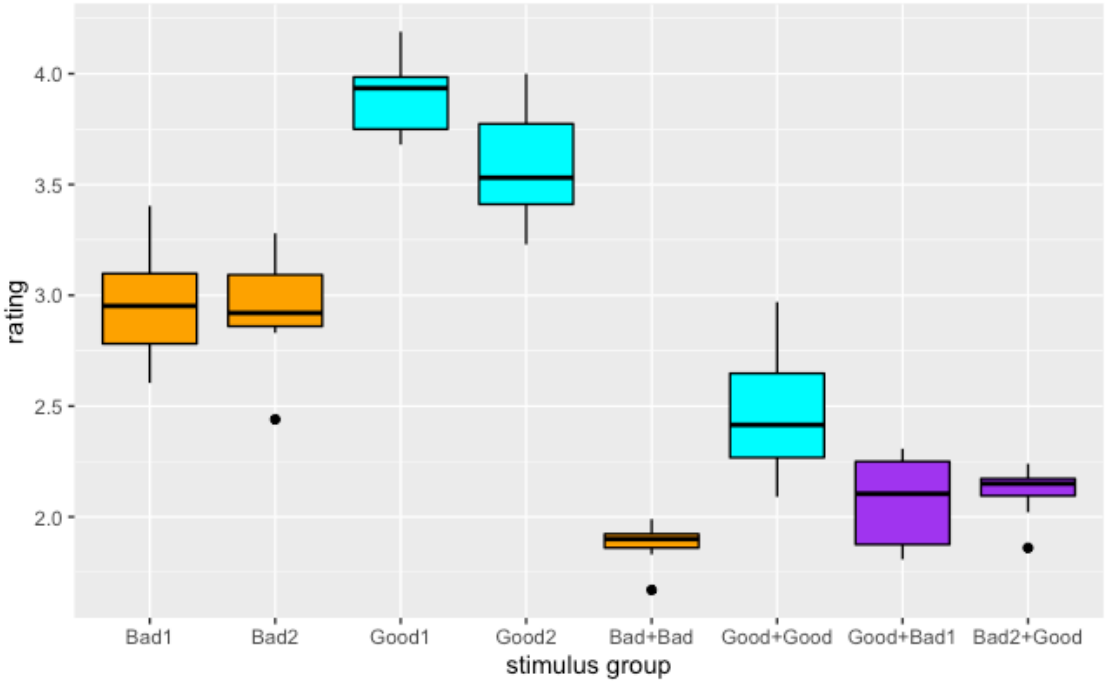


Figure 8: Box plot of Amelio stimulus group mean ratings

Table 1: Mean rating for each Amelio stimulus group

Group	Mean
Bad1	2.97
Bad2	2.93
Good1	3.90
Good2	3.60
Bad+Bad	1.88
Good+Good	2.48
Good+Bad1	2.07
Bad1+Good	2.11

The following sections describe the statistical analysis performed to test both the Amelioration and Length Hypotheses. All statistical testing described in these sections was carried out using the R programming language version 3.6.0 (R Core Team 2019).

5.5.1 Amelioration Hypothesis

The primary hypothesis to be tested, the Amelioration Hypothesis, states that a word known to be rated as poorly acceptable can see its rating improve if an additional sequence of segments that has separately been rated as more acceptable is prepended or appended to the poorly acceptable word.

To test this hypothesis, two ordinal regression models with mixed effects, A1 and A2, were fitted using the `clmm()` function of the `ordinal` package (version 2019.12.10) (Christensen 2019). The stimulus groups included in the models were as follows: Bad1, Bad2, Good+Bad1, and Bad2+Good. Both models have an am treatment-coded binary predictor representing whether a stimulus is a member of one of the groups where amelioration has been attempted (Good+Bad1 and Bad2+Good). Additionally, each model has a second binary predictor: $b1$ or $b2$ (for A1 and A2, respectively). These predictors indicate whether a stimulus contains a Bad1 or Bad2 stimulus as a substring. The purpose of fitting two models in this fashion that merely reverse the binary values of a predictor was simply the ease of obtaining significance values for the am predictor against different reference category stimulus groups. Table 2 shows the predictor coding scheme, and Figure 9 shows the fitted models in R syntax:

Table 2: Amelioration Hypothesis predictor coding

	am	b1	b2
Bad1	0	1	0
Bad2	0	0	1
Good+Bad1	1	1	0
Bad2+Good	1	0	1

```
rating ~ 1 + am + b1 + am*b1 + (1|ppt)
rating ~ 1 + am + b2 + am*b2 + (1|ppt)
```

Figure 9: Amelioration Hypothesis models A1 and A2

The models include $rating$ (an integer from 1 to 5, inclusive) as the ordinal response variable, as well as the binary predictors, and a term for their interaction. Random intercepts are also included for participant (ppt).

Tables 3-6 show the estimated predictor coefficients for each model along with the threshold coefficients computed by each model to determine the boundaries between the ordinal response values.

Table 3: Amelioration Hypothesis model A1 coefficients

	Est	Std. Err.	z val	Pr(>abs(z))
am1	-1.65	0.10	-16.75	<2e-16
b11	0.08	0.09	0.88	0.38
am1:b11	-0.21	0.14	-1.58	0.11

Table 4: Amelioration Hypothesis model A1 threshold coefficients

	Est	Std. Err.	z val
1-2	-2.26	0.17	-13.46
2-3	-0.81	0.16	-4.97
3-4	0.97	0.16	5.90
4-5	3.05	0.18	17.09

Table 5: Amelioration Hypothesis model A2 coefficients

	Est	Std. Err.	z val	Pr(>abs(z))
am1	-1.86	0.10	-18.60	<2e-16
b21	-0.08	0.09	-0.88	0.38
am1:b21	0.21	0.14	1.58	0.11

The A1 and A2 models have the Bad2 and Bad1 stimulus groups as their respective reference categories. Tables 3 and 5 show that in both models, the am predictor has a significant effect. I.e., the differences between Bad1 and Good+Bad1 (model A2) and Bad2 and Bad2+Good (model A1) are both significant. For both models, however, the estimated coefficient has a negative sign, indicating that the effect on stimulus rating of "amelioration" (i.e., concatenation) is in fact significantly negative. Thus, although the null hypothesis, that amelioration has no effect on acceptability ratings, can be rejected, it is clear that the Amelioration Hypothesis cannot be accepted.

The same can be demonstrated by an examination of probabilities. As seen in Table 1, the unameliorated Bad stimulus groups have mean ratings of very close to 3. Thus, a reasonable test

Table 6: Amelioration Hypothesis model A2 threshold coefficients

	Est	Std. Err.	z val
1-2	-2.34	0.17	-13.91
2-3	-0.89	0.16	-5.46
3-4	0.88	0.16	5.40
4-5	2.97	0.18	16.66

of the Amelioration Hypothesis would be to compare the probability, as estimated by the model, of stimuli in the ameliorated groups (Good+Bad1 and Bad2+Good) having a rating of greater than 3 (i.e., 4 or 5) with the same probability for the unameliorated Bad stimuli. If the stimuli in the ameliorated groups were found more likely to receive a rating of greater than 3 than the unameliorated stimuli, then it could be said that the results obtained were in keeping with the Amelioration Hypothesis.

Using the coefficients in Tables 5 and 6 (for model A2), it is possible to compute the log odds of having a rating of less than or equal to 3 for all possible combinations of values for the predictors am and $b2$. Table 7 shows these values.

Table 7: Log odds of rating ≤ 3

	am=0	am=1	Row mean
b2=0	0.88	2.74	1.81
b2=1	0.97	2.61	1.79
Col mean	0.93	2.68	

For easier interpretation, Table 8 shows these log odds converted to probabilities, which were then subtracted from 1 to yield $p(rating > 3)$:

Table 8: Probability of rating > 3

	am=0	am=1	Row mean
b2=0	0.29	0.06	0.18
b2=1	0.28	0.07	0.17
Col mean	0.28	0.07	

As seen in the "Col mean" row of Table 8, the unameliorated Bad stimuli ($am = 0$) have a probability of having a rating greater than 3 roughly four times greater than the stimuli in the

ameliorated groups ($am = 1$). This finding runs counter to the prediction of the Amelioration Hypothesis, allowing it to be rejected.

These models can also be used to compute the exact probability of having a specific star rating for each of the four stimulus groups. The rating with the highest probability can then be taken as the predicted rating for stimuli in each group. These predictions can be compared with the actual most frequent rating observed in the data. Table 9 shows these predictions (made using model A2), along with the actual most frequent rating, shown in parentheses in each cell.

Table 9: Model-predicted ratings and actual most-frequent ratings

	am=0	am=1
b2=0	3 (3)	1 (1)
b2=1	3 (3)	1 (1)

5.5.2 Length Hypothesis

The secondary hypothesis to be tested, the Length Hypothesis, states that increasing word length in general will have a negative effect on acceptability ratings.

To test this hypothesis, two separate mixed effects ordinal regression models, L1 and L2, were again fitted, this time to the data in the Bad1, Bad2, Good1, and Good2 stimulus groups. Both models have a $long$ binary predictor representing whether a stimulus has one or two syllables. Additionally, each model has a second binary predictor: $good$ or bad (for L1 and L2, respectively). As their names indicate, these predictors determine whether a stimulus is Good or Bad. Table 10 shows the predictor coding scheme, and Figure 10 shows the fitted models in R syntax:

```
rating ~ 1 + good + long + good*long + (1|ppt)
rating ~ 1 + bad + long + bad*long + (1|ppt)
```

Figure 10: Length Hypothesis models L1 and L2

These models again include $rating$ as the ordinal response variable, as well as the two

Table 10: Length Hypothesis predictor coding

	good	bad	long
Bad1	0	1	0
Bad2	0	1	1
Good1	1	0	0
Good2	1	0	1

binary predictors, and a term for their interaction. Random intercepts are also included for participant (ppt).

Tables 11-14 show the estimated predictor coefficients for each model and threshold coefficients computed to determine the boundaries between the ordinal response values.

Table 11: Length Hypothesis model L1 coefficients

	Est	Std. Err.	z val	Pr(>abs(z))
good1	1.84	0.10	18.78	<2e-16
long1	-0.07	0.09	-0.78	0.44
good1:long1	-0.60	0.13	-4.56	5.25e-06

Table 12: Length Hypothesis model L1 threshold coefficients

	Est	Std. Err.	z val
1-2	-2.47	0.14	-17.82
2-3	-0.83	0.13	-6.59
3-4	0.86	0.13	6.88
4-5	2.67	0.13	29.00

The Bad1 stimulus group is the reference group for model L1. From Table 11, it can be seen that the $long$ predictor on its own is not significant in this model (i.e., the difference between groups Bad1 and Bad2 is not significant). Model L2 has the Good1 group as its reference group, and Table 13 shows that, contrary to the results for model L1, the effect of $long$ is significant (i.e., the difference between groups Good1 and Good2 is significant). The estimated coefficient has a negative sign, indicating that the model is more likely to predict lower ratings for Good2 stimuli than Good1 stimuli. Thus, the null hypothesis, that length has no effect on acceptability ratings, can be rejected, but only for Good stimuli.

Table 13: Length Hypothesis model L2 coefficients

	Est	Std. Err.	z val	Pr(>abs(z))
bad1	-1.84	0.10	18.77	<2e-16
long1	-0.67	0.09	-7.08	1.45e-12
bad1:long1	0.60	0.13	4.55	5.25e-06

Table 14: Length Hypothesis model L2 threshold coefficients

	Est	Std. Err.	z val
1-2	-4.30	0.15	-28.66
2-3	-2.67	0.14	-19.79
3-4	-0.98	0.13	-7.63
4-5	0.83	0.13	6.54

This apparent discrepancy can also be illustrated through comparison of estimated probabilities, in a similar fashion to the analysis for the Amelioration Hypothesis. However, in this case, separate tests are necessary for the Good and Bad stimuli groups, as the means of groups are far apart. Here, the probabilities of Good1 and Good2 stimuli having ratings of less than or equal to 4 (which is close to the mean for Good1) will be compared, as well as the probabilities of Bad1 and Bad2 stimuli having ratings less than or equal to 3 (near the mean for Bad1).

Table 15: Probability of Good stimuli having a rating ≤ 4

	good=1
long=0	0.70
long=1	0.82

As seen in Table 15, disyllabic Good stimuli (group Good2) have a roughly 17% greater probability of having a rating of less than 4 than monosyllabic Good stimuli (group Good1).

As seen in Table 16, as with the comparison between the Good groups, the disyllabic stimuli (group Bad2) have a greater probability of having a lower rating (less than 4) than the monosyllabic stimuli (group Bad1). However, the difference is much smaller, with the longer stimuli only having a roughly 3% greater probability of a lower rating.

As with the Amelioration Hypothesis models, these models can also be used to predict star ratings for stimuli in each group, which can then be compared with the actual most frequent rating

Table 16: Probability of Bad stimuli having a rating ≤ 3

	good=0
long=0	0.70
long=1	0.72

observed in the data. Table 17 shows these predictions (made using model L1), along with the actual most frequent rating, shown in parentheses in each cell.

Table 17: Model-predicted ratings and actual most-frequent ratings

	good=0	good=1
long=0	3 (3)	4 (5)
long=1	3 (3)	4 (4)

5.5.3 Correlation with reference model predictions

Pearson's r was computed using R's `cor()` function to test the correlation between mean stimulus ratings across participants and grammaticality scores computed by the MaxEnt and AvBg reference phonotactic models. The mean ratings and AvBg scores had a correlation of $r = 0.44$, and the mean ratings and MaxEnt scores had a correlation of $r = -0.71$. The moderately strong negative correlation (recall that greater MaxEnt scores mean lower grammaticality) with the MaxEnt scores is supportive of the MaxEnt model's approach to grammaticality, where the impact of phonotactic violations in one part of a word cannot be mitigated by other parts of a word with fewer or no violations. The correlation with the AvBg scores is rather weak, suggesting less utility as a predictor of acceptability.

6 Discussion

6.1 Amelioration Hypothesis

Unsurprisingly given the breakdown of mean ratings shown in Figure 8, the models fitted to test the Amelioration Hypothesis reveal no evidence to support the hypothesis. However, considering the often unnatural sound of the concatenated trisyllabic stimuli, there remains a significant possibility that the ratings of both groups of "ameliorated" trisyllables (Good+Bad1 and Bad2+Good) were driven downward by this factor. Such an effect, if strong enough, could have potentially buried any visible evidence of amelioration. Section 7 discusses the potential for future work to address this possibility.

6.2 Length Hypothesis

The models fitted to test the Length Hypothesis do reveal the presence of a detrimental effect of increased word length on acceptability. Notably, the effect is only significant in its influence on the ratings of Good stimuli.

One might suspect that this discrepancy was due to structural differences between the Good and Bad stimuli. A structural difference is in fact present, but along the "length" dimension rather than that of "goodness". The feature of being one or two syllables in length does not occur with equal frequency across the Good and Bad stimuli groups. Although each participant was only exposed to eight stimuli in each of the four groups included in the model for the length hypothesis (Bad1, Bad2, Good1, Good2), the use of two conditions doubled the number of stimuli in the Bad1 and Good2 groups. For this reason, the Bad stimuli as a whole include twice as many monosyllables as disyllables, and the Good stimuli twice as many disyllables as monosyllables. In other words, monosyllabic words account for two-thirds of Bad stimuli, but only a third of Good stimuli.

So in fact there are important overall structural differences between the Good and Bad stimuli. Good stimuli are mostly disyllabic, and Bad stimuli are mostly monosyllabic. The length

effect may appear more potent with respect to the Good stimuli simply because of the fact that this group contains so many more “long” stimuli than the Bad group.

7 Possibilities for future research

The most obvious next step in continuing this line of research would be to tackle the issue of effect on ratings of audio splicing. One possibility would be to perform an experiment (“Son of Amelio”) wherein participants would compare trisyllabic stimuli constructed using the concatenation process described in Section 5.2.1 with trisyllabic stimuli of similar expected acceptability (perhaps the identical words) recorded as a single unit. If some bias against stimuli with audio spliced in the manner discussed in this paper could be detected and quantified, it might be possible to “un-bias” the ratings collected in Amelio and determine whether the results are affected.

Another potential avenue of interest would be to score the Amelio stimuli with other phonotactic models of note, and see how they fare as rating predictors, especially against the MaxEnt model.

APPENDIX A: LISTS OF STIMULI

Tables 18-26 list the stimuli used in Amelio in each major group, including orthography, pronunciations in IPA format, and condition where applicable.

Table 18: Bad1 stimuli

Orthography	IPA	Condition
blylve	blaɪlv	A
cheeln	tʃiln	B
gwoib	gwɔɪb	A
gwoog	gwug	B
meerzh	miɹʒ	A
noilp	nɔɪlp	B
pwud	pwʌd	A
shmoin	ʃmɔɪn	B
sphutch	sfʌtʃ	A
thaict	θeɪkt	B
thwev	θwɛv	A
toisp	tɔɪsp	B
yerlth	jɝlθ	A
yermp	jɝmp	B
zheelb	ʒilb	A
zlerd	zlɝd	B

Table 19: Good2 stimuli

Orthography	IPA	Condition
binkaw	bɪŋ'kɔ	A
binzy	'bɪnzi	B
claffy	'klæfi	A
custaw	kə'stɔ	B
duffroo	də'fɹu	A
frola	'fɹoʊlə	B
kinday	kɪn'deɪ	A
linyoo	lɪn'ju	B
paibry	'peɪbɹi	A
pilsha	'pɪlʃə	B
scutoo	skə'tu	A
siskoo	sɪ'sku	B
spobby	'spɑbi	A
styba	'staɪbə	B
voopny	'vupni	A
wistaw	wɪ'stɔ	B

Table 20: Good+Bad1 stimuli

Orthography	IPA	Condition
binkawthwev	bɪŋ'kɔθwɛv	B
binzytoisp	'bɪnzi.tɔɪsp	A
claffymeerzh	'klæfi.miɹʒ	B
custawzlerd	kə'stɔzlɝd	A
duffrooyerlth	də'fɹujɝlθ	B
frolayermp	'fɹoʊlə.jɝmp	A
kindaypwud	kɪn'deɪpwʌd	B
linyoocheeln	lɪn'jutʃiln	A
paibrysphutch	'peɪbɹi.sfʌtʃ	B
pilshagwoog	'pɪlʃə.gwug	A
scutooblylve	skə'tublaɪlv	B
siskoothaict	sɪ'skuθeɪkt	A
spobbyzheelb	'spɑbi.ʒilb	B
stybanoilp	'staɪbə.nɔɪlp	A
voopnygwoib	'vupni.gwɔɪb	B
wistawshmoin	wɪ'stɔʃmɔɪn	A

Table 21: Good1 stimuli

Orthography	IPA
clill	klɪl
dind	dɪnd
floist	flɔɪst
gruld	gɹʌld
preel	pɹil
quex	kwɛks
sonce	sɑns
stoon	stun

Table 22: Bad2 stimuli

Orthography	IPA
beesya	'bisjə
dwoufy	'dwaʊfi
nifthay	nɪf'θeɪ
pustigh	pə'staɪ
sibbja	'sɪbdʒə
thuchaw	θə'tʃɔ
yaisfa	'jeɪsfə
zhifmoo	ʒɪf'mu

Table 23: Good+Good stimuli

Orthography	IPA
bontapreesh	'bɑntəˌpɹiʃ
drellymup	'dɹɛliˌmʌp
fuscaydrask	fə'skeɪdɹæsk
luspygronk	'lʌspiˌgɹɔŋk
queelascose	'kwiləˌskoʊs
skerdaclinn	'skɝdəˌklɪn
trullayswadge	tɹə'leɪswædʒ
yoondysterce	'jundiˌstɝs

Table 24: Bad+Bad stimuli

Orthography	IPA
buspahshoamf	bə'spɑʃoʊmf
dwuthaypwibe	dwə'θeɪpwaɪb
nuvzightwadh	nəv'zaɪtwæð
shilnighjaimf	ʃɪl'naɪdʒeɪmf
thaizbyzloit	'θeɪzbiˌzlɔɪt
wubbjygwoash	'wʌbdʒiˌgwoʊʃ
yivvawzerln	jɪ'vɔzɝln
zidnoythailve	zɪd'nɔɪθeɪlv

Table 25: Bad2+Good stimuli

Orthography	IPA
chuffthoogress	tʃəf'θugɹɛs
jicrayskack	dʒɪ'kɹeɪskæk
pikwighdreen	pɪ'kwaɪdɹin
shydwacront	'ʃaɪdwəˌkɹɑnt
thishtayqueeld	θɪʃ'teɪkwild
wudnahprack	wə'dnɑpɹæk
yupmoystite	jəp'mɔɪstaɪt
zutfowtrell	zə'tfaʊtɹɛl

Table 26: Warm-up stimuli

Orthography	IPA
dwouf	dwaʊf
jufftha	'dʒʌfθə
perndy	'pɝndi
spest	spɛst

APPENDIX B: ORTHOGRAPHY

Orthography was chosen for maximal consistency, as the written forms shown to participants were intended more as a guide to the sounds in the audio than official "spelling forms". However, some concessions were made toward familiarity to existing English forms where it seemed helpful.

Vowels

Most vowels had multiple possible written forms, often conditioned by whether the vowel occurred word-internally vs. word-finally or on a concatenation boundary.

- [ɑ]: <o> 'spobby', <ah> concatenated 'buspahshoamf'
- [aɪ]: <y> 'styba', <igh> word-final or concatenated 'pustigh, nuvzightwadh', <i> 'yupmoystite'
- [aʊ]: <ou> 'dwoufy', <ow> concatenated 'zutfowtrell'
- [æ]: <a> 'claffy'
- [ʌ], [ə]: <u> 'pwud', <a> word-final or concatenated 'pilsha, frolayermp'
- [eɪ]: <ai> 'thaict', <ay> word-final or concatenated 'kinday, fuscaydrask'
- [ɛ]: <e> 'thwev'
- [ɝ]: <er> 'yerlth'
- [i]: <ee> 'cheeln', <y> word-final or concatenated 'binzy, luspygronk'
- [ɪ]: <i> 'binkaw'
- [oʊ]: <o> 'frola', <oa> 'buspahshoamf'
- [ɔɪ]: <oi> 'gwoib', <oy> concatenated 'zidnoythailve'
- [ɔ]: <aw> 'binkaw', <o> 'luspygronk'
- [u]: <oo> 'gwoog'

Consonants

Consonants were often doubled after short vowels; e.g.:

- [æf]: ‘claffy’
- [əf]: ‘duffroo’
- [ɪl]: ‘clill’
- [ɛl]: ‘drellymup’

Other accommodations were made to familiar English spelling conventions:

- [lv#]: <lve> ‘blylve’
- [tʃ#]: <tch> ‘sphutch’
- [#sf]: <sph> ‘sphutch’
- [ŋk]: <nk> ‘binkaw’
- [ns#]: <nce> ‘sonce’
- [ædʒ#]: <adge> ‘trullayswadge’
- [oʊs#]: <ose> ‘queelascose’
- [ɜ˞s#]: <erce> ‘yoondysterce’
- [aɪb#]: <ibe> ‘dwuthaypwibe’
- [aɪt#]: <ite> ‘yupmoystite’
- [#kw]: <qu> ‘quex’
- [ks#]: <x> ‘quex’

[k] was often represented as <c>, when it seemed more natural:

- ‘claffy’
- ‘thaict’

- ‘fuscaydrask’

[ʒ] and [ð] lack unambiguous English spellings, and were thus always consistently represented using <zh> and <dh>, respectively, as in ’meerzh’ and ’nuvzightwadh’.

BIBLIOGRAPHY

Albright, A. (2008). Gradient phonological acceptability as a grammatical effect. *Phonology. https://doi. org/10.1017 S.*

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90*(2), 119-161.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The CELEX lexical database (cd-rom).

Boersma, Paul & Weenink, David (2021). *Praat: doing phonetics by computer* [Computer program]. URL http://www.praat.org/

The CMU Pronouncing Dictionary. Carnegie Mellon University. URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Christensen, R. H. B. (2019). *ordinal - Regression Models for Ordinal Data.* R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Coetzee, A. W., & Pater, J. (2011). 13 The Place of Variation in Phonological Theory. *The handbook of phonological theory*, 401.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *arXiv preprint cmp-lg/9707017.*

Daland, R. (2015). Long words in maximum entropy phonotactic grammars. *Phonology, 32*(3), 353-383. doi:http://dx.doi.org/10.1017/S0952675715000251

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*(3), 379-440.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR), 33*(1), 31-88.

Pater, J. (2015, June 2). *Wellformedness = probability?* UMass Amherst Computational Phonology Lab Blog. http://blogs.umass.edu/comphon/2015/06/02/wellformedness-probability/

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schutze, C. T. (2017). *The Empirical Base of Linguistics*. Freie Universitat Berlin, Universitatsbibliothek.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and speech, 40*(1), 47-62.

Wilson, C., & Capodieci, F. (2008, June). *Phonotactic learning program*. URL https://linguistics.ucla.edu/people/hayes/Phonotactics/

Winter, B. (2020). *Statistics for linguists: An introduction using R.*